Ethical and Responsible AI as a Means of Diagnosing and Eliminating Bias

Barbara J. Thompson[1], Ayris A. Narock[1,2], Daniel E. da Silva[1,3], Alexa J. Halford[1], Burcu Kosar[1,4], Mykhaylo S. Shumko[1,3]

1. NASA/GSFC Center for HelioAnalytics
2. ADNET Systems, Inc.
3. University of Maryland
4. Catholic University of America

Abstract

AI and Machine Learning (ML) are powerful tools that allow us to analyze and derive knowledge from information that may not be accessible using more traditional methods. AI/ML opens the gateway to exploring more complex relationships in both our data and in our practices as a community of scientists. However, the results of any ML model must be examined to ensure that they are valid and beneficial; otherwise the practitioners may act on false or misleading results. The NASA Framework for the Ethical Use of AI identifies principles and practices that are fundamental to Ethical AI.

There are many ways that Ethical AI can be leveraged to improve equity and fairness in our field. For example, there are practices in machine learning that can be used to clearly diagnose the factors behind human decisions, allowing us to pinpoint the presence and sources of bias. This can become a roadmap for ensuring fairness in future decisions. This presentation will review strategies for implementing Ethical/Responsible AI and discuss how they can be used to advance Diversity, Equity and Inclusion.

Ethical AI is more than a code of behavior. It forms the foundation of AI-enabled scientific progress in general.

What is Ethical and Responsible AI?

- "Ethical AI is artificial intelligence that adheres to well-defined ethical guidelines regarding fundamental values, including such things as individual rights, privacy, non-discrimination, and non-manipulation. Ethical AI places fundamental importance on ethical considerations in determining legitimate and illegitimate uses of AI."
- "Organizations that apply ethical AI have clearly stated policies and well-defined review processes to ensure adherence to these guidelines."

Excerpts from C3.ai glossary – https://c3.ai/glossary/artificial-intelligence/ethical-ai/

What is Ethical and Responsible AI?

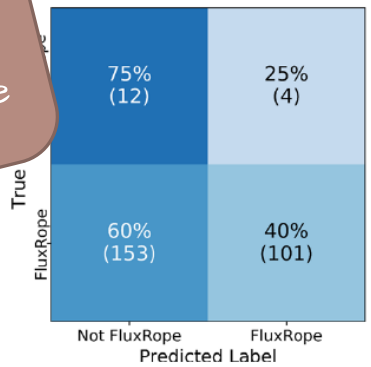
- European Commission: *Ethics guidelines for trustworthy AI* - April 2019
  https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- DOD Adopts Ethical Principles for Artificial Intelligence - February 24, 2020
  https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/
- Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government - December 3, 2020
  https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/
- NASA: *Framework for the Ethical Use of Artificial Intelligence (AI)* - April 2021
  https://ntrs.nasa.gov/citations/20210012886
- AGU workshops underway to establish principles and responsibilities for AI/ML Ethics in the Earth, space, and environmental sciences

NASA Framework for the Ethical Use of Artificial Intelligence (AI)

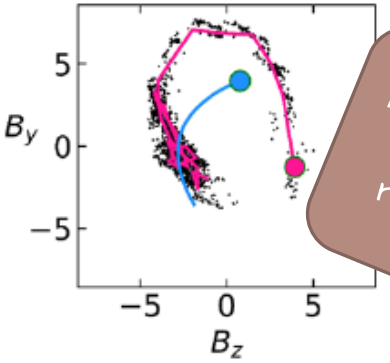
Fair	AI systems must include considerations of how to treat people, including scrubbing solutions to mitigate discrimination and bias, preventing covert manipulation, and supporting diversity and inclusion.
Explainable and Transparent	Solutions must clearly state if, when, and how an AI system is involved, and AI logic and decisions must be explainable. AI solutions must protect intellectual property and include risk management in their construction and use. AI systems must be documented.
Accountable	Organizations and individuals must be accountable for the systems they create, and organizations must implement AI governance structures to provide oversight.
Secure and Safe	AI systems must respect privacy and do no harm. Humans must monitor and guide machine learning processes. AI system risk tradeoffs must be considered when determining benefit of use.
Human-Centric and Societally Beneficial	AI systems must obey human legal systems and must provide benefits to society. At the current state of AI humans must remain in charge, though future advancements may cause reconsideration of this requirement.
Scientifically and Technically Robust	AI systems must adhere to the scientific method NASA applies to all problems, be informed by scientific theory and data, robustly tested in implementation, well-documented, and peer reviewed in the scientific community.

Considerations for Greater Explainability / Interpretability

"This model is biased towards predicting the negative case."



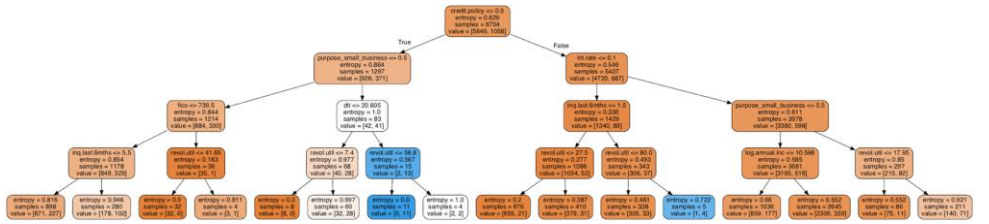
Techniques can give insight
Globally [performance over the entire se
or
Locally [prediction in a specific instance

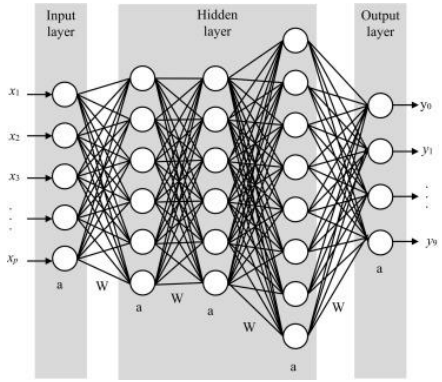


"This case is misclassified because it is not well represented in the training data."

Examine training data for bias / patterns:

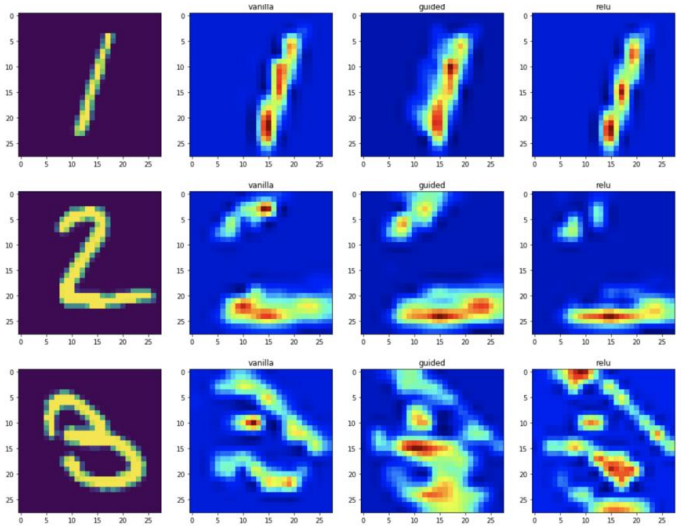
Choose simplest method available:



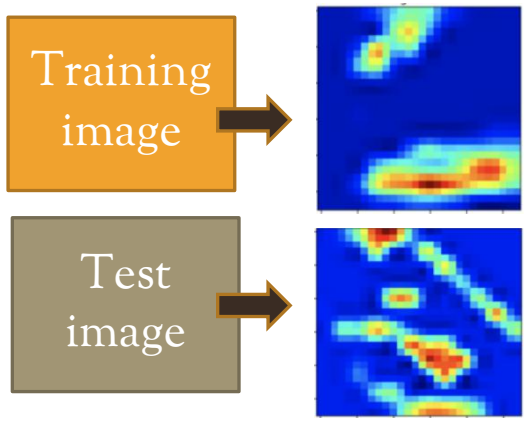


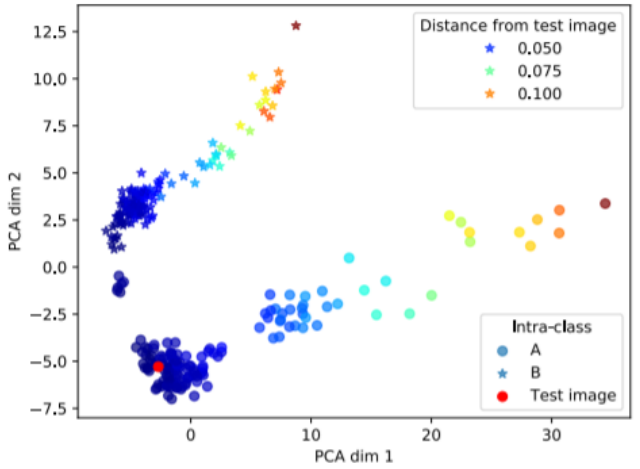
Sample Techniques for Greater Explainability / Interpretability

Saliency Mapping

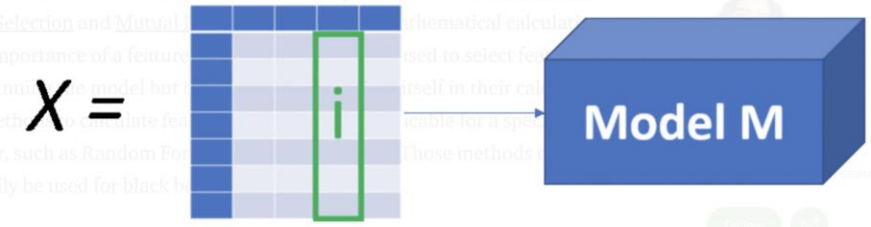


Activation Analysis





Feature Sensitivity Analysis



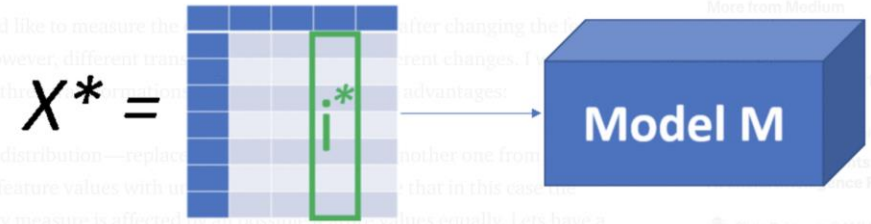


Image source: https://medium.com/@elhac93027/understand-your-black-box-model-using-sensitivity-analysis-practical-guide-ef6ac4175e55

Using Ethical AI to diagnose and mitigate bias

Fair	**All AI-based decisions and systems should include in their design goals:** 1) fair treatment of people, 2) elimination of discrimination and bias, 3) prevention of covert manipulation 4) support of diversity and inclusion.
Explainable and Transparent	**Explainability is a pathway to identify and eliminate bias.** It is difficult to ensure that human decisions are not influenced by biased factors because it is impossible to isolate the process. This is possible with an algorithm, however - if the results are fully explainable and transparent, it will be clear whether biases had an influence.
Accountable	**Practitioner assumes responsibility for all decisions made on behalf of an entity**, whether the decision be provided by a human or an algorithm.
Secure and Safe	**A thorough consideration of the security implications for your team and the model's results.** This includes those *impacted* by a result or a decision, personal information used in data, and members of the design team that produced the model.
Human-Centric and Societally Beneficial	**An AI decision or a model's results should include a full assessment of potential impacts (and adopt mitigation strategies where relevant) prior to release.** A result that is statistically true may not be societally beneficial, or may have some impacts that can be detrimental.
Scientifically and Technically Robust	**Failure to adhere to scientific standards and community review can be damaging to science and humanity.** A well-intentioned effort that adopts all of the above principles can still have detrimental impacts if the foundation and performance of the model itself is flawed or insufficiently characterized.

So how do we make use of Ethical and Responsible AI to identify and eliminate bias?

Explainable and responsible machine learning can be used to clearly diagnose the factors behind human decisions, allowing us to pinpoint the presence and sources of bias.

Even if a human does not want to be influenced by factors such as race, gender, age, it is impossible to remove that data from a human mind once it is present.

However, an AI can explicitly be told to exclude e.g. race, gender, age from computation. Even more importantly, an *AI can be trained to reproduce human decisions*, and determine which input features that had an influence on a decision.

So how do we make use of Ethical and Responsible AI to identify and eliminate bias?

When you review a proposal, a paper, or a job application;

When you write a recommendation or reply to an inquiry;

When you choose collaborators, team members, or coauthors:

The complexity of human decisions makes it impossible to fully quantify the influence of prejudices and biases.

By training a computer to evaluate and reproduce human decisions, we can determine which factors impact performance accuracy. If the model requires e.g. age, gender, nationality to reproduce a decision, it allows us to identify these connections and makes it possible to improve the decision process.

However, most models can only identify relationships in the data; we are responsible for a full diagnosis of implications and correlation vs. cause.

What are the sources of bias in AI systems?

Bias can appear at any stage of an AI system's development:

- It can appear in the formation and structure of the development team
- Bias can influence decisions to support or implement projects
- Bias can be introduced in the data used in the project
- The manipulation or processing of the data can include technical and societal biases
- The selection and training of the model may be inappropriate for the input population
- The human interpretation of [technically correct] results can introduce unconscious and/or erroneous associations
- The deployment or distribution of the AI can cause or reinforce societal biases

Types of Bias to be aware of in AI systems

Selection Bias	We introduce selection bias into our data when our choice of data sources is skewed in such a way that the proper randomization is not achieved. When undetected, practitioners end up analyzing and modeling samples that are not representative of the population.
Self-Selection Bias	Self-selection bias is a different form of selection bias; It arises in any situation in which the examples we use to build our dataset select themselves into a group, for example, any data source that was created by volunteers.
Omitted-Variable Bias	Omitted-variable bias occurs when a dataset leaves out one or more features that are necessary for an accurate prediction. As a result, our model will attribute the effect of the missing features to the features that were present during training.
Supporter/Funder Bias	Funding bias is the tendency of a scientific study to support the interest of the study's sponsor or originator. This can influence the data sample collection and skew results by only investigating models that are likely to be beneficial to the supporter.
Sampling Bias	Sampling bias occurs when the data used to train a model doesn't reflect the distribution of the samples that the model will receive whilst in production — this is probably one of the most common types of biases observed in real-world scenarios because accurate (and easily utilized) training data often is a limited subset of all available data.
Stereotype Bias	A stereotype is simply a widely held but fixed and oversimplified image or idea of a particular type of person or a thing. Stereotype bias is common in human-produced samples, such as Books, Photo Archives, Social Media, Online Forums and comments.

Source: Kurtis Pykes, https://towardsdatascience.com/tackling-different-types-of-bias-in-data-projects-29e326660639

Types of Bias to be aware of in AI systems (continued)

Systematic Value Distortion	Systematic value distortion usually occurs when the device or process being used to record measurements has a problem — our machine learning algorithm would make suboptimal predictions when the model is deployed into production.
Experimenter Bias	Experimenter bias is when the researcher unconsciously affects the results. data. or a participant in an experiment due to their prior beliefs or hypothesis. In a machine learning context, model builders may unconsciously process the data in a manner that confirms their preexisting beliefs or hypothesis. Additionally, a model builder may continually increase the training iterations until the model outputs a result that is in alignment with their hypothesis.
Labeling Bias	Labeled data are the most commonly used data in AI and machine learning because of the relatively straightforward goal of having the model reproduce the labels. However, when a biased person or process assigns labels to unlabeled data this introduces labeling bias. Practitioners are predisposed to use clearly labeled data for their models, and can make the erroneous assumption that the labels had a clear correspondence to the system they are modeling.

Source: Kurtis Pykes, https://towardsdatascience.com/tackling-different-types-of-bias-in-data-projects-29e326660639

Turning "AI Fails" into successes

Example: Automated review of job applications

[This is a hypothetical scenario based partially on a true story]

A large company had enough application data to train an AI to predict which applicants had the highest chance of being hired. The training data was based on the decision of humans who reviewed the applications.

The hope is that the company could review the applications more systematically and eliminate inconsistencies in the human-based evaluations.

However, the AI learned that performance accuracy improved when it excluded candidates based on data associated with gender, age, race, and nationality.

Turning "AI Fails" into successes

Example: Automated review of job applications

The company realized that the goal of *consistent* evaluation of application of resumes is not the same as *fair* evaluation.

Bias was introduced in the input sample (the population of applicants was imbalanced relative to the general population), the labeling (humans unwittingly or intentionally included factors in their decision that should not have been used), and the model goal (in order to optimize performance, it reinforced and perpetuated bias).

However, there were major benefits to this exercise: 1) clear evidence of bias in hiring practices 2) clearer identification of how and where these biases influenced decisions 3) a means of mitigating the source of biases, and 4) a pathway to more equitable decisions, a more diverse and talent-based workforce, and greater worker satisfaction overall.

Conclusions

The principles of Ethical AI are not unique to AI and machine learning models.

They are general principles that are relevant to our research, the way we do science, and the way we interact with each other and society. These principles make us better scientists, and better human beings.

Models and AI are often viewed with suspicion and skepticism relative to human decisions. However, human decisions are not reproducible, and even a sincere effort to understand the factors behind our behavior can be inaccurate and unreliable.

AI is not a panacea to eliminate bias. However, a deliberate responsible effort to incorporate the beneficial aspects of AI and data science can create a pathway towards greater equity and inclusion.